

# A Genome-Based Species Taxonomy of the *Lactobacillus* Genus Complex

 Stijn Wittouck,<sup>a,b</sup>  Sander Wuyts,<sup>a</sup>  Conor J. Meehan,<sup>c,d</sup>  Vera van Noort,<sup>b</sup>  Sarah Lebeer<sup>a</sup>

<sup>a</sup>Research Group Environmental Ecology and Applied Microbiology, Department of Bioscience Engineering, University of Antwerp, Antwerp, Belgium

<sup>b</sup>Centre of Microbial and Plant Genetics, KU Leuven, Leuven, Belgium

<sup>c</sup>School of Chemistry and Bioscience, University of Bradford, Bradford, United Kingdom

<sup>d</sup>BCCM/ITM Mycobacterial Culture Collection, Institute of Tropical Medicine, Antwerp, Belgium

**ABSTRACT** There are more than 200 published species within the *Lactobacillus* genus complex (LGC), the majority of which have sequenced type strain genomes available. Although genome-based species delimitation cutoffs are accepted as the gold standard by the community, these are seldom actually checked for new or already published species. In addition, the availability of genome data is revealing inconsistencies in the species-level classification of many strains. We constructed a *de novo* species taxonomy for the LGC based on 2,459 publicly available genomes, using a 94% core nucleotide identity cutoff. We reconciled these *de novo* species with published species and subspecies names by (i) identifying genomes of type strains and (ii) comparing 16S rRNA genes of the genomes with 16S rRNA genes of type strains. We found that genomes within the LGC could be divided into 239 *de novo* species that were discontinuous and exclusive. Comparison of these *de novo* species to published species led to the identification of nine sets of published species that can be merged and one species that can be split. Further, we found at least eight *de novo* species that constitute new, unpublished species. Finally, we reclassified 74 genomes on the species level and identified for the first time the species of 98 genomes. Overall, the current state of LGC species taxonomy is largely consistent with genome-based species delimitation cutoffs. There are, however, exceptions that should be resolved to evolve toward a taxonomy where species share a consistent diversity in terms of sequence divergence.

**IMPORTANCE** The *Lactobacillus* genus complex is a group of bacteria that constitutes an important source of strains with medical and food applications. The number of bacterial whole-genome sequences available for this taxon has been increasing rapidly in recent years. Despite this wealth of information, the species within this group are still largely defined by older techniques. Here, we constructed a completely new species-level taxonomy for the *Lactobacillus* genus complex based on ~2,500 whole-genome sequences. As a result of this effort, we found that many genomes are not classified to their correct species, and we were able to correct these. In addition, we found that some published species are abnormally large, while others are too small. Finally, we discovered at least eight completely novel species that have not been published before. Our work will help the field to evolve toward a more meaningful and complete taxonomy, based on whole-genome sequences.

**KEYWORDS** genomics, *Lactobacillus*, taxonomy, species delimitation, core genome


Since the advent of whole-genome sequencing, it has become clear that current bacterial taxonomy is not always consistent with bacterial evolutionary history. In particular, some official taxa are not monophyletic, and taxa of the same rank often differ in the diversity they represent in terms of sequence identity or shared gene

**Citation** Wittouck S, Wuyts S, Meehan CJ, van Noort V, Lebeer S. 2019. A genome-based species taxonomy of the *Lactobacillus* genus complex. *mSystems* 4:e00264-19. <https://doi.org/10.1128/mSystems.00264-19>.

**Editor** Sean M. Gibbons, Institute for Systems Biology

**Copyright** © 2019 Wittouck et al. This is an open-access article distributed under the terms of the [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/).

Address correspondence to Sarah Lebeer, [sarah.lebeer@uantwerpen.be](mailto:sarah.lebeer@uantwerpen.be).

 Comparative genomics of the *Lactobacillus* Genus Complex using ~2,500 genomes reveals that species are discontinuous and exclusive, that some species should be merged or split and that many genomes are currently not classified to their closest type strain.

**Received** 24 April 2019

**Accepted** 12 August 2019

**Published** 3 September 2019

content. The genus *Lactobacillus* is a prime example of both problems (1). First, it is larger than a typical bacterial family. Second, the genus is paraphyletic: the genera *Pediococcus*, *Leuconostoc*, *Weissella*, *Oenococcus*, and *Fructobacillus* are also descendants of the most recent common ancestor of *Lactobacillus*. Together, these six genera form a monophyletic taxon that is sometimes referred to as the *Lactobacillus* genus complex (LGC) (2). The LGC is an important source of bacteria with medical, food, and feed applications (1) and is a key player in the human microbiome (3, 4). Because of all this, LGC taxonomy has been the topic of much study and debate, especially during recent years (1, 5–7).

In an attempt to systematically correct inconsistencies in the official taxonomy of all *Bacteria* and *Archaea*, Parks et al. (8) recently constructed the Genome Taxonomy Database (GTDB). They built a phylogeny of a dereplicated subset of all sequenced bacterial and archaeal genomes and determined the diversity within each official taxon using a heuristic called relative evolutionary divergence (RED). With this information, they corrected taxa that were not monophyletic or that contained a diversity much above or below the average diversity of their rank. In their alternative, genome-based taxonomy, all six LGC genera were integrated into the family *Lactobacillaceae*, and the genus *Lactobacillus* was split into 15 smaller genera.

The approach of Parks et al. (8) to split/merge only taxa that are outliers in terms of diversity within their rank is appealing for the taxonomic ranks from phylum to genus, because these are relatively arbitrary and have no meaningful cutoffs in terms of diversity. The species level, however, is different. First, while there is agreement that the ranks from phylum to genus are arbitrary, there exists discussion on the possible biological meanings of the species rank. For example, one recent study found that there exists a relatively large gap between within-species and between-species genome distances; in other words, that the “genome space” shows a discontinuity corresponding to the species level (9). If this would prove to be indeed the case, it would mean that a prokaryotic species concept based on this discontinuity would hold real biological meaning. In another recent study, the hypothesis that taxa of the species rank show the property of exclusivity more than taxa of other ranks was explored (10). A taxon is exclusive if all of its members are more related to each other than to anything outside of the taxon. The authors concluded that their data falsified this hypothesis: many taxa, from many different ranks, indeed show the property of exclusivity, but that this was not more often the case for taxa of the species rank. Thus, there is currently no agreement on whether a particular biological significance should be attached to the species rank. However, a second way in which the species rank stands out is more pragmatic. In contrast to the other taxonomic ranks (e.g., the genus rank), relatively fixed similarity cutoffs to separate species are generally accepted and largely seem to correspond to the historically defined species (9, 11).

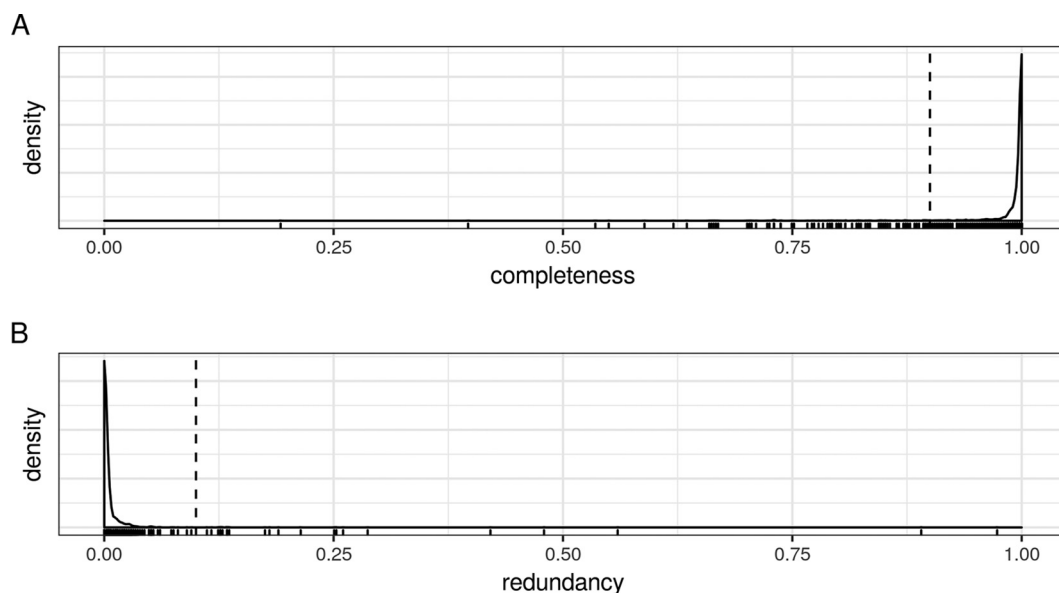
The cutoff that is commonly used to separate bacterial species based on their genome sequences, is 94 to 96% average nucleotide identity (ANI) (11, 12). Despite this cutoff being accepted as the gold standard for bacterial species delimitation, it is seldom checked for type strain genomes of validly published species and subspecies. Type strain genomes of different species are sometimes more closely related than ~95% ANI, while type genomes of subspecies of the same species are sometimes more distantly related than ~95% ANI. For example, it has been shown that the type strains of *Lactobacillus casei* and *Lactobacillus zeae* are too closely related to consider them separate species (13). This has led to the rejection of the *L. zeae* name. A related problem is that newly sequenced genomes, of type strains or otherwise, are not systematically checked for similarity against type strains. In the NCBI assembly database, uploaders of new assemblies are free to choose which species name they assign to the genomes. This has resulted in many classification inconsistencies. For example, we have previously shown that many genomes annotated as *Lactobacillus casei* in the NCBI database are in reality more closely related to the *Lactobacillus paracasei* type strain (14). Finally, it is possible that some sequenced genomes are so distantly related to all currently described species that they should be considered a new species.

In this work, we aimed to infer *de novo* species within the LGC by downloading all genomes belonging to this taxon from the NCBI assembly database and clustering them based on pairwise genome similarities. Advantages of this genome-based, *de novo* species taxonomy are as follows. (i) Future work on comparing LGC species based on genomes can be performed with a complete and nonredundant data set of exactly one representative genome per species. (ii) Comparative genomics studies of an individual species can proceed with a complete set of all genomes that are within species range of its type strain. Furthermore, we reconciled our *de novo* species with the currently established LGC species by identifying the genomes of type strains and by comparing 16S rRNA gene sequences to those of type strains. On the basis of this reconciliation, we propose changes to the official taxonomy in the form of splits or mergers of species. As a cornerstone for genome quality control, pairwise genome comparisons, and phylogeny inference, we made use of single-copy core genes (SCGs) of the LGC. The classical approach to identify SCGs in a set of genomes is to first determine the complete pangenome of the genomes, but this process is very computationally expensive, especially for data sets spanning more than a single species. Therefore, we developed a novel approach that should be easily applicable to any (large) genome data set. Briefly, we first identify candidate SCGs in a small, random subset of genomes and then extract those candidates from all genomes in the data set using HMM profiles. Our pipeline for rapid SCG extraction was implemented in progenomics, a toolkit for prokaryotic comparative genomics, and is available at <https://github.com/SWittouck/progenomics>.

## RESULTS

**SCGs and genome quality.** We downloaded all 2,558 genomes that belong to the *Lactobacillaceae* and *Leuconostocaceae* families from GenBank (see Table S1 in the supplemental material) and identified single-copy core genes (SCGs) in this data set. Our strategy for SCG extraction consisted of four steps: (i) full gene family clustering on a random subset of 30 genomes, (ii) selection of candidate SCGs from these gene families using a very soft single-copy presence threshold, (iii) search for the candidate SCGs in all genomes, and (iv) filtering of the definitive SCGs from the candidates using a fixed single-copy presence threshold. After visual inspection of the single-copy presence of the candidate SCGs in the full genome data set (see Fig. S1 in the supplemental material), we retained candidate SCGs with >95% single-copy presence. This strategy resulted in 411 SCGs *sensu lato* ("*sensu lato*" because we did not enforce 100% single-copy presence). We performed quality control of all genomes based on this set of SCGs (Fig. 1). More specifically, we determined two quality measures for each genome: completeness, defined here as the percentage of *sensu lato* SCGs that are present in the genome, and redundancy, defined as the percentage of *sensu lato* SCGs showing two or more copies in the genome. The large majority of genomes had a completeness close to one and redundancy close to zero. Enforcing a minimum completeness of 90% and maximum redundancy of 10% resulted in 2,459 high-quality genomes, or 96.1% of all genomes.

Some of the genomes with low estimated completeness belonged to the genus *Sharpea*, which is part of the family *Lactobacillaceae* according to the NCBI taxonomy but seems to be phylogenetically quite distant to it (8). Other low-completeness genomes were flagged as "incomplete" in the NCBI database or were derived from metagenome data. Interestingly, some genomes belonging to *Lactobacillus iners*, the LGC species with the smallest genome, had very high completeness values. This suggests that genome reduction did not (significantly) influence our completeness estimates. As an extra check, we compared our completeness and redundancy estimates to the estimates available in the Genome Taxonomy Database (GTDB), where possible. There was a good correspondence between both estimates, with the exception of a small number of genomes (among which were the aforementioned *Sharpea* genomes) that had low completeness estimates by our pipeline, but high estimates in



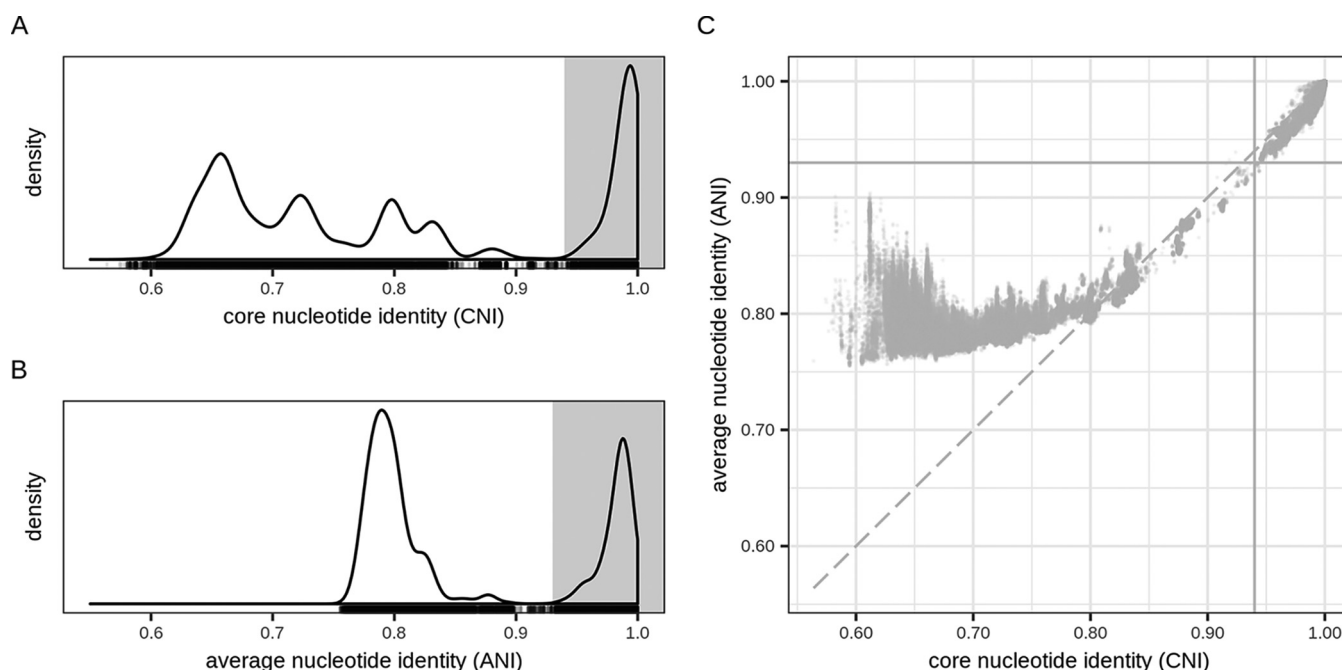
**FIG 1** Genome quality control based on single-copy core gene (SCG) completeness and redundancy. (A) Density of genome completeness. For each genome, the percentage of SCGs that were present is shown. (B) Density of genome redundancy. For each genome, the percentage of SCGs with one or more extra copies is shown. The small vertical bars at the bottom of the graphs represent individual genomes and are shown to visualize outliers more clearly.

the GTDB (Table S1). This was not unexpected, since our pipeline uses LGC-specific core genes for quality estimation, while the GTDB uses a set of taxon-specific core genes.

**Pairwise genome similarities and clustering.** We compared every genome to every other genome using the core nucleotide identity (CNI), which we calculated from the supermatrix of SCGs. While average nucleotide identity (ANI) is the most commonly used genome similarity measure, it suffers from conceptual and practical drawbacks. First, since it is based on all genomic regions that are orthologous between two genomes, these regions will be different for each pairwise comparison. Thus, the type and amount of information are different across genome comparisons. Second, ANI is slow to compute for large data sets because it scales quadratically with the number of input genomes. The recently introduced fastANI (9) is orders of magnitude faster, but it suffers from the same conceptual problem and remains an approximation. Thanks to our novel approach for the rapid extraction of SCGs, computing CNI similarities becomes feasible in almost linear time.

The density of these pairwise CNI values was strongly bimodal (Fig. 2A): there was one sharp peak with CNI values of  $>94\%$ , and a flatter peak (including multiple subpeaks) with values of  $<90\%$ . This indicated that two genomes were either very closely related (high CNI) or relatively distantly related (low CNI). Therefore, genomes were clustered into *de novo* species by grouping together genomes with similarities equal to or higher than  $94\%$ , using single-linkage clustering. This resulted in 239 genome clusters.

In single-linkage clustering, an object is added to a cluster when it exceeds a similarity threshold to at least one object within that cluster (as opposed to, for example, an average similarity threshold used in average linkage clustering). One problem that can arise with this strategy is that individual similarities within clusters can potentially be low, much lower than the clustering cutoff used. To assess whether this problem occurred for our clusters, for each cluster we calculated the minimum CNI between genomes within the cluster and the maximum CNI between genomes in the cluster and genomes outside of the cluster. Almost all within-cluster CNI values were higher than  $94\%$  (Fig. S2). There was only one exception: two genomes in cluster 231 had a CNI of  $93.6\%$ , but even this value was very close to the species boundary. Furthermore, all 239 genome clusters met the criterion of exclusivity: all their within-



**FIG 2** Pairwise genome distance values between LGC genomes. (A) All pairwise core nucleotide identity (CNI) similarities between LGC genomes. The gray area indicates the CNI values of  $>94\%$  (the same-species range). (B) All pairwise fastANI similarities between LGC genomes. The gray area indicates the fastANI values of  $>93\%$  (the hypothetical same-species range). (C) CNI versus ANI values. Only genome pairs with fastANI values of  $>75\%$  are shown, since the fastANI tool does not compute values under 75%.

cluster CNI values were higher than their between-cluster CNI values. This included cluster 231 (later identified as *Lactobacillus sakei*): its highest between-cluster CNI value was 81.6% (to cluster 221, later identified as *Lactobacillus curvatus*).

In addition to the CNI, we also calculated a second similarity measure between all our genomes: the fastANI similarity, which is a fast approximation of ANI. Just like the CNI distribution, the fastANI distribution was strongly bimodal (Fig. 2B). Its high-value peak covered the range 93% to 100%. A scatterplot of the CNI and fastANI similarities (Fig. 2C) showed that both measures were strongly correlated for closely related genomes (around or above the species cutoff) and that the fastANI values were generally lower than the CNIs in this range. For more distantly related genomes, with CNI of  $<80\%$ , the values were clearly less correlated, and the fastANI values were (much) higher than their CNI counterparts. The CNI species delimitation cutoff of 94% roughly corresponded to a fastANI cutoff of 93%. At these cutoffs, both similarity measures were in strong agreement on whether two genomes belonged to the same species or not.

To assess whether the species level held properties absent from the other taxonomic ranks, we calculated the exclusivity of clusters at various CNI cutoff values. The results showed that for CNI clustering cutoffs between 88% and 96%, all clusters were exclusive (Fig. S3). This was not the case for most cutoffs outside of this range.

**Comparison with published species.** We then attempted to label all 239 genome clusters (*de novo* species) with the name of a published species. We did this in two stages: (i) identification of genomes belonging to type strains of published (sub)species (further called “type genomes”) and (ii) for the remaining clusters, extraction of 16S rRNA genes from the genomes and comparison to 16S rRNA genes from type strains of published (sub)species (further called “type 16S genes”).

In the first cluster-naming stage, we identified type genomes present in our data set (Table S2). First, we cross-referenced the strain names that were listed in the NCBI assembly database with type strain names of validly published (sub)species found on the List of Prokaryotic Names with Standing in Nomenclature (LPSN) (15, 16) or



**TABLE 1** Inconsistencies between published and *de novo* species

Cluster	Merger or split	Type strain(s) present	Minimum CNI
135	Merger	<i>L. amylophilus</i> , <i>L. amylophilus</i>	1
218	Merger	<i>P. acidilactici</i> , <i>P. lolii</i>	0.981
179	Merger	<i>L. fructivorans</i> , <i>L. homohiochii</i>	0.979
178	Merger	<i>W. jogaejeotgali</i> , <i>W. thailandensis</i>	0.974
173	Merger	<i>L. micheneri</i> , <i>L. timberlakei</i> , <i>L. koso</i>	0.964
238	Merger	<i>L. casei</i> subsp. <i>casei</i> , <i>L. zeae</i>	0.954
236	Merger	<i>Leuc. mesenteroides</i> subsp. <i>mesenteroides</i> , <i>Leuc. mesenteroides</i> subsp. <i>cremoris</i> , <i>Leuc. mesenteroides</i> subsp. <i>dextranicum</i> , <i>Leuc. mesenteroides</i> subsp. <i>jonggajibkimchii</i> , <i>Leuc. suionicum</i>	0.952
225	Merger	<i>Leuc. gelidum</i> subsp. <i>gelidum</i> , <i>Leuc. inhae</i> , <i>Leuc. gasicomitatum</i>	0.946
233	Merger	<i>L. gasseri</i> , <i>L. paragasseri</i>	0.946
141	Merger	<i>L. kimchii</i> , <i>L. bobalius</i>	0.944
64	Split	<i>L. aviarius</i> subsp. <i>araffinosus</i>	1
149	Split	<i>L. aviarius</i> subsp. <i>aviarius</i>	0.953

Prokaryotic Nomenclature Up-to-date (PNU) (17). This resulted in the identification of 367 type genomes within our data set, belonging to 226 unique names (species or subspecies) from 211 species. Several type genomes for the same name were often found because many type strains have been sequenced multiple times. Next, we supplemented this type genome list by manually scanning the literature for extra type genomes, including also species that were not published validly (i.e., not in an official taxonomic journal). This yielded type genomes for an additional 11 species. There were different reasons why these manually identified type genomes had not been detected by our automated approach. Some of them were not represented on LPSN and PNU because they were not validly published. For other type genomes, their species were validly published, but too recently for them to be already on LPSN or PNU. Then there were some special cases. *Lactobacillus nuruki* (18) was a validly published name but was not listed on LPSN or PNU, probably because its genus name was misspelled as “Lactobacilus” in the NCBI database and in the paper. *Convivina intestini* was missed by our automated approach because we did not include this genus in our search, while it does appear to be part of the LGC, more specifically of the family *Leuconostocaceae* (16). The *Lactobacillus musae* type genome was missed because its type strain name (strain 313) consists only of numbers and was consequently filtered out by our pipeline. Finally, *Leuconostoc lactis* had a type genome in the NCBI database that did not pass our quality control, but we could also identify a type genome of *Leuconostoc argentinum*, a later synonym of *Leuconostoc lactis* (*Leuc. lactis*). This type genome had not been identified by our automated pipeline because it filtered out rejected (sub)species names.

Of our set of 239 genome clusters, 210 contained one or more type genomes (Table S1). The large majority of these clusters (200 of them) contained type genomes from exactly one species and could thus be given the name of that species. Ten clusters contained type genomes from two or three different species, indicating that the type strains of these species were more closely related than the 94% CNI cutoff (Table 1). These clusters were cluster 135 with *Lactobacillus amylophilus* and *Lactobacillus amylophilus*, cluster 141 with *Lactobacillus kimchii* and *Lactobacillus bobalius*, cluster 173 with *Lactobacillus micheneri*, *Lactobacillus timberlakei*, and *Lactobacillus koso*, cluster 178 with *Weissella jogaejeotgali* and *Weissella thailandensis*, cluster 179 with *Lactobacillus fructivorans* and *Lactobacillus homohiochii*, cluster 218 with *Pediococcus acidilactici* and *Pediococcus lolii*, cluster 225 with *Leuconostoc gelidum*, *Leuconostoc inhae*, and *Leuconostoc gasicomitatum*, cluster 233 with *Lactobacillus gasseri* and *Lactobacillus paragasseri*, cluster 236 with *Leuconostoc suionicum* and *Leuconostoc mesenteroides*, and cluster 238 with *Lactobacillus casei* and *Lactobacillus zeae*. We assigned the name of the species that was described first to these clusters. In one situation, two subspecies of the same species had a type genome in different genome clusters: *Lactobacillus aviarius* subsp. *aviarius* and *Lactobacillus aviarius* subsp. *araffinosus*. We named the genome cluster with the latter type strain *Lactobacillus araffinosus*.

**TABLE 2** Genome clusters without a type strain genome<sup>a</sup>

Cluster	NCBI species	No. of 16S rRNA genes	16S hits	Species <sup>b</sup>
155	<i>L. backii</i>	28	<i>L. backii</i> , <i>L. iwatensis</i>	<i>L. backii</i> (?)
185	<i>L. bombi</i>	6	<i>L. bombi</i>	<i>L. bombi</i> (?)
4	<i>L. coleohominis</i>	1	<i>L. coleohominis</i>	<i>L. coleohominis</i> (?)
175	<i>L. kefir</i>	12	<i>L. kefir</i>	<i>L. kefir</i> (?)
165	<i>L. panisapium</i>	1	<i>L. panisapium</i>	<i>L. panisapium</i> (?)
7	<i>Leuc. carnosum</i>	4	<i>Leuc. gelidum</i> subsp. <i>aenigmaticum</i> , <i>Leuc. carnosum</i> , <i>Leuc. rapi</i>	<i>Leuc. carnosum</i> (?)
157	<i>Leuc. citreum</i>	32	<i>Leuc. citreum</i> , <i>Leuc. palmae</i> , <i>Leuc. lactis</i> , <i>Leuc. holzapfelii</i> , <i>Leuc. gelidum</i> subsp. <i>aenigmaticum</i>	<i>Leuc. citreum</i> (?)
176	<i>Leuc. pseudomesenteroides</i>	25	<i>Leuc. gelidum</i> subsp. <i>aenigmaticum</i> , <i>Leuc. pseudomesenteroides</i> , <i>Leuc. rapi</i> , <i>Leuc. holzapfelii</i>	<i>Leuc. pseudomesenteroides</i> (?)
129	<i>P. parvulus</i>	1	<i>P. parvulus</i>	<i>P. parvulus</i> (?)
12	<i>W. ceti</i>	25	<i>W. ceti</i>	<i>W. ceti</i> (?)
184	<i>W. cibaria</i>	93	<i>W. cibaria</i>	<i>W. cibaria</i> (?)
11	<i>W. hellenica</i>	1	<i>W. hellenica</i> , <i>W. bombi</i>	<i>W. hellenica</i> (?)
167	<i>W. koreensis</i>	10	<i>W. koreensis</i>	<i>W. koreensis</i> (?)
164	NA	1	NA	New species 1
166	NA	1	NA	New species 2
168	NA	5	NA	New species 3
169	NA	5	NA	New species 4
192	NA	5	NA	New species 5
211	NA	1	NA	New species 6
25	NA	1	NA	New species 7
3	NA	1	NA	New species 8
206	<i>W. bombi</i>	0	NA	Unidentified species 1
207	<i>W. hellenica</i>	0	NA	Unidentified species 2
1	NA	0	NA	Unidentified species 3
132	NA	10	<i>L. fornicalis</i>	Unidentified species 4
190	NA	4	<i>L. musae</i>	Unidentified species 5
196	NA	4	<i>L. panisapium</i>	Unidentified species 6
2	NA	5	<i>L. cerevisiae</i> , <i>L. yonginensis</i>	Unidentified species 7
217	NA	1	<i>L. caviae</i>	Unidentified species 8

<sup>a</sup>*Leuc.*, *Leuconostoc*; NA, not available.<sup>b</sup>A question mark in parentheses indicates that this species is a “best guess” (see “Comparison with published genomes” in Results).

In the second cluster-naming stage, we manually named the 29 remaining genome clusters in which we had not found type genomes (Table 2). We could assign a species name to 13 genome clusters with relative certainty because two sources of information were in agreement: (i) the NCBI species labels attached to the genomes in the cluster and (ii) matches to our in-house-constructed database of 16S rRNA genes of validly published species not yet used to name a cluster via type genomes. For these genome clusters, we included a question mark in the species name to indicate that these names are not 100% certain but rather a “best guess.” Interestingly, we labeled eight clusters as “new species” because they yielded 16S rRNA gene sequences that showed no hits to our type 16S rRNA gene database, indicating that they were different from all known and validly published species. Each of these clusters consisted of exactly one genome. Finally, for the remaining nine clusters, the available information did not allow us to make a “best guess” nor to consider them novel species with certainty. We labeled those clusters “unidentified species.” An overview of all named genome clusters can be found in Table S3.

We inferred a maximum likelihood species tree of the LGC using one representative genome for each of the 239 genome clusters (Fig. 3). For representative genomes, we took the genomes with the largest number of SCGs, because the type strain genomes were not always of the best quality and because some genome clusters did not contain type strains. In this phylogeny, the eight new species and nine unidentified species were also included. Two of the unidentified species were situated in the *Weissella* genus; the other new and unidentified species could be found in the *Lactobacillus* genus. Some of the new/unidentified species had long “tip branches” in the tree, indicating that they were relatively far removed from the other species. We annotated



the tree with the phylogroups defined by Zheng et al. (6) and predicted the lifestyle of the new species using the work of Duar et al. (2). Strikingly, the isolation source of each of the eight new species confirmed their predicted lifestyles, whether they were vertebrate-adapted, insect-adapted, or free-living bacterial species (Table 3). For nomadic phylogroups, this was trivially the case, since they can occur in any environment.

msystems.asm.org 8



**TABLE 3** Isolation sources, phylogroups, and predicted lifestyles of new species

Species	Isolation source (NCBI)	Phylogroup	Phylogroup lifestyle
New species 1	A <i>Marmota</i> species	<i>L. delbrueckii</i> group	Vertebrate-adapted
New species 2	Urine catheter	<i>L. reuteri</i> group	Vertebrate-adapted
New species 3	Kimchi	<i>L. plantarum</i> group	Nomadic
New species 4	Kimchi	<i>L. plantarum</i> group	Nomadic
New species 5	Gut of <i>Bombus ignitus</i>	<i>L. kunkei</i> group	Insect-adapted
New species 6	Cow rumen	<i>L. salivarius</i> group	Vertebrate-adapted
New species 7	<i>Apis dorsata</i>	<i>L. kunkei</i> group	Insect-adapted
New species 8	Human gut	<i>L. delbrueckii</i> group	Vertebrate-adapted

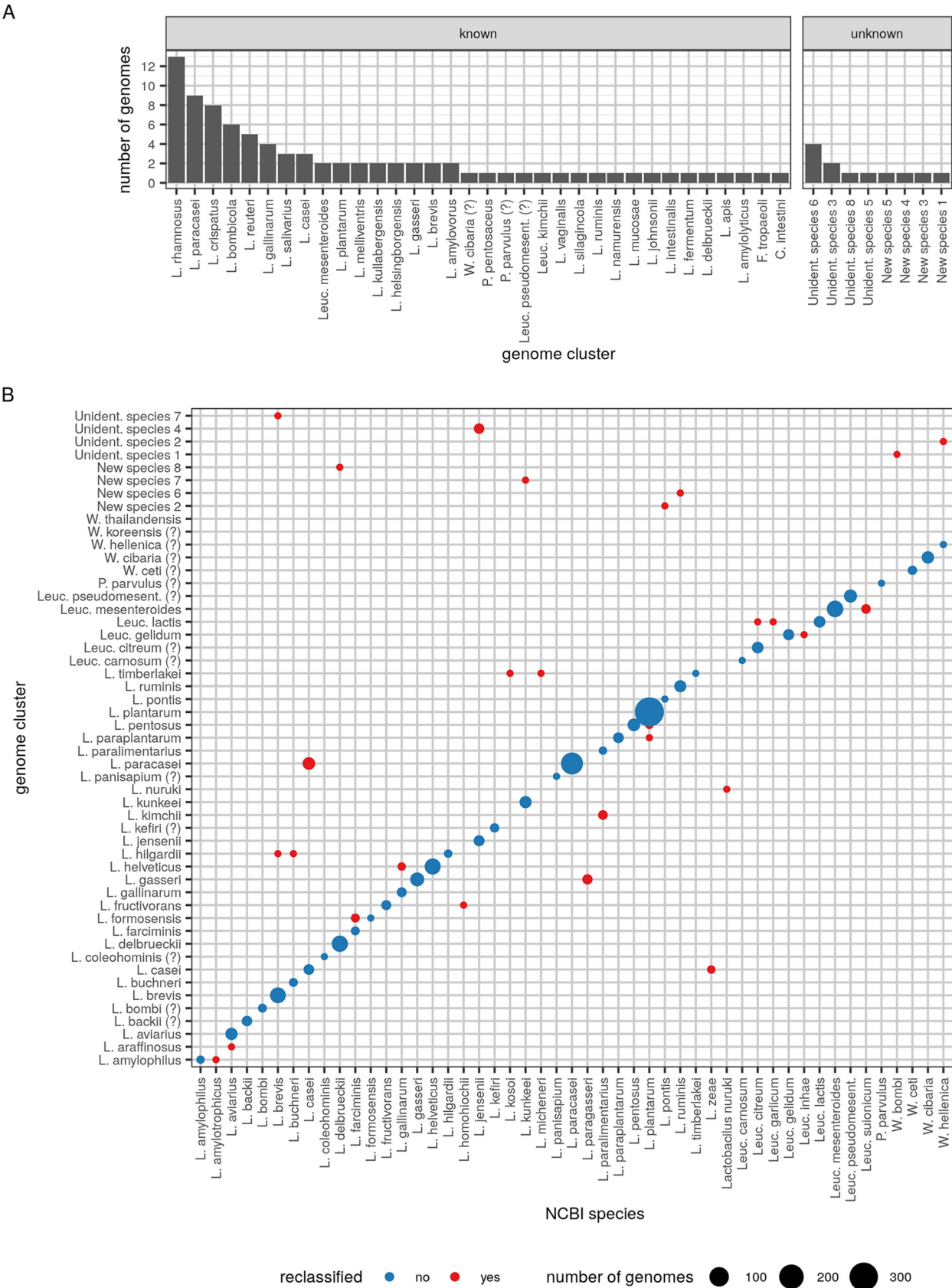
genome clustering and cluster naming pipelines, these genomes were automatically assigned to a species: either an existing one or a new or unidentified species (Fig. 4A). The most frequently identified species in this group was *Lactobacillus rhamnosus*, with 13 previously unclassified genomes assigned to it. Interestingly, several unclassified genomes belonged to clusters that we labeled as new or unidentified species.

Of the genomes that did have a species label in the NCBI database, 74 were reclassified to other species by our approach (Fig. 4B). Of these reclassifications, 21 were a direct consequence of the species mergers and splits we described. Next, 38 reclassifications were clear cases of genomes whose NCBI species label did not correspond to their most closely related type strain. We reclassified 20 genomes from *Lactobacillus casei* to *Lactobacillus paracasei*, 5 genomes from *Lactobacillus paralimentarius* to *Lactobacillus kimchii*, 3 genomes from *Lactobacillus farciminis* to *Lactobacillus formosensis*, 2 genomes from *Lactobacillus gallinarum* to *Lactobacillus helveticus*, 2 genomes from *Lactobacillus plantarum* to *Lactobacillus pentosus*, 1 genome from *Lactobacillus brevis* to *Lactobacillus hilgardii*, 1 genome from *Lactobacillus buchneri* to *Lactobacillus hilgardii*, 1 genome from *Lactobacillus plantarum* to *Lactobacillus paraplantarum*, and 1 genome from *Leuconostoc citreum* to *Leuc. lactis*. Finally, 15 genomes with a species label in the NCBI database ended up in genome clusters considered by us as new or unidentified species.

## DISCUSSION

We developed a novel approach to extract single-copy core genes (SCGs) from large genome data sets and applied it to calculate pairwise CNI as well as fastANI distances between all high-quality genomes within the LGC. We observed that for relatively closely related genomes (CNI of 80% to 100%), the fastANI similarities were lower than the CNI similarities. This could be explained by the fact that the CNI is based exclusively on conserved genes, and these genes can be expected to mutate more slowly. In contrast, fastANI is based on all homologous genome regions between two genomes and thus encompasses also recently gained, potentially fast-evolving genes. For more distantly related genomes (CNI < 80%), fastANIs were higher than CNIs. A possible explanation for this is that for these genome pairs, it becomes more difficult for an ANI calculation tool to detect the homologous regions that show lower sequence identity; the fastANI distance might thus be based only on the homologous regions with higher sequence identity because they are easier to detect. Thus, the CNI might be a better reflection of evolutionary distance because it is based on a fixed set of (prealigned) genes.

We observed a discontinuity in the density of pairwise CNI values, confirming observations made by others (e.g., reference 9). This could be taken as evidence that the species level holds real biological meaning. Alternatively, one could argue that the observed discontinuity is the result of genome sampling bias. Indeed, there have been many comparative genomics studies in recent years where the aim was to compare as many strains as possible from the same species (e.g., references 19 and 20), and this might have biased genome databases toward sets of closely related strains. However, the observation that only the CNI range of 88% to 96% shows full exclusivity of clusters



**FIG 4** LGC genome reclassifications. (A) Classification of genomes that are currently unclassified in NCBI database, using the CNI-based genome clusters. (B) Reclassification of genomes that have an NCBI species label available but were found in a different CNI species cluster. Species names that are attached to an identical set of genomes in the NCBI and CNI classifications are not shown.

is a stronger confirmation of the biological meaning of the species rank. This should be tested in more taxa to further explore whether the property of exclusivity is connected to the rank of species and to study the bacterial species question in general.

In discussing the biological significance of the species rank, we focused on species properties defined in terms of genome-genome distances. However, a number of bacterial species concepts in terms of the underlying, generative biological processes have also been proposed. In the species concept of Dykhuizen and Green (21), bacterial species are groups of strains within which recombination is common, while recombination between species is comparatively rare. Another popular species concept is the ecotype concept of Cohan and Kane (22), which states that bacterial species occupy a given ecological niche. Within a species, an innovative mutation is able to purge diversity through a vertical sweep, but this process cannot occur between different species. It is not yet clear at this point what the potential discontinuity and/or exclusivity of bacterial species would mean for the validity of the various bacterial species concepts that have been put forward.

Using a pragmatic, genome distance-based species concept, we clustered genomes of the LGC into *de novo* species using single-linkage clustering with a threshold of 94% CNI. By following this strategy, we implicitly defined a species as a set of strains for which it holds that each strain shows at least 94% core genome sequence identity to at least one other strain in the species and not to a single strain outside of the species. Using this species definition, we found that species membership in the LGC was almost transitive: all strains within a species were more than 94% similar to each other (there was only one exception to this). This has a number of interesting consequences. First, it means that, when using this species concept, species are very robust to strain/genome sampling; adding extra strains will not lead to mergers of species that were previously separated. Second, it means that for strain classification, we can calculate the similarity of the new strain to one (random) representative strain per species. When the strain is >94% similar for its core genes to one strain of a species, it will also be >94% similar to all other strains in the species. Similarly, when the strain is <94% similar to one strain of a species, it will also be <94% similar to all other strains in the species. Finally, transitivity means that clustering with single linkage or complete linkage will yield identical results. Therefore, transitivity is a very useful property to have for a given species definition and genome space. However, it should be noted that despite the fact that only one exception to transitivity was found in this work, more exceptions might be found when additional genome sequences become available in the future.

We identified situations where type strain genomes from more than one published species were found in the same *de novo* species. We argue that in these situations, the published species names should be merged. In taxonomic terms, such a merger means that the species names are considered heterotypic synonyms, i.e., names associated with different type strains but referring to the same species. Because the 94% CNI cutoff value we used (corresponding to approximately 93% fastANI) was lower than what is commonly suggested (11, 12), more species mergers than splits can be expected *a priori*. This cutoff value was chosen because it corresponded to the discontinuity we observed in the density of pairwise CNIs. A consequence of this is that for some of the mergers suggested by our pipeline, the type strain genomes will still be relatively dissimilar. Whether a merger is still preferable in these “gray area” situations depends on the species concept one wishes to use and more specifically, on the importance one attributes to the observed discontinuity in genome space for taxonomy.

Four genome clusters contained type strain genomes from different species that were very closely related (CNI > 96%). *Lactobacillus micheneri* and *Lactobacillus timberlakei* belonged to a cluster with a minimum CNI of 96.4%. These species were described in the same recent publication (23), so it is not straightforward which name should be considered the earlier synonym. *Weissella thailandensis* (24) and *Weissella jogaejeotgali* (25) were present in a cluster with a minimum CNI of 97.4%. The *W. jogaejeotgali* paper mentioned that its type strain was closely related to *W. thailandensis*, with 99.39% 16S identity. According to our genome-based comparison, *W.*

*jogaejeotgali* should be considered a later heterotypic synonym. The type strains of *Lactobacillus fructivorans* (26) and *Lactobacillus homohiochii* (27) were present in a cluster with a minimum CNI of 97.9%. It was already known that these type strains are very closely related (28). In addition, it has been shown that other strains classified as *L. homohiochii* clearly form a separate species from *L. fructivorans* (28). Thus, in this case, the *L. homohiochii* type strain seems badly chosen. According to the International Code of Nomenclature of Prokaryotes (29), this means that *L. homohiochii* should be considered a later heterotypic synonym of *L. fructivorans* and a new type strain should be proposed for the other strains currently classified as *L. homohiochii*. Unfortunately, none of these other strains have been sequenced yet. Finally, a similar situation occurred for *Pediococcus acidilactici* (30) and *Pediococcus lolii* (31), which were in the same cluster with a minimum CNI of 98.1%. It has already been found by Wieme et al. that the *P. lolii* type strain DSM 19927 is an *P. acidilactici* strain (32). *P. lolii* should be considered a later synonym of *P. acidilactici* and if other *P. lolii* strains cluster separately, a new type strain and species name should be proposed for them.

Some other cases of potential species mergers were less clear-cut. Before the species *Lactobacillus amylophilus* was introduced (33), with type strain DSM 20534, its type strain had been classified to *Lactobacillus amylophilus*, with type strain DSM 20533 (34). We now found that these two type strain genomes were almost identical (CNI of ~100%) at the level of their core gene sequences. However, the publication that introduced the species *L. amylophilus* (33) clearly shows that strain DSM 20534 is relatively distant from strain DSM 20533 based on a comparison of *pheS* and *rpoA* gene sequences. Thus, we suspect that a mistake was made and that both genomes are actually the same strain, while the other type strain has not yet been sequenced. It is difficult to guess which of the type strains was actually sequenced (twice), but luckily, strain DSM 20533 was also sequenced by another institute. This second DSM 20533 genome was also identical to the other two. Therefore, it is highly likely that all three genomes are DSM 20533, the *L. amylophilus* type strain, and that the *L. amylophilus* type strain has not been sequenced yet.

In a 2012 paper (35), it was suggested that *Lactobacillus kimchii* (36) and *Lactobacillus bobalius* (37) be considered later heterotypic synonyms of *Lactobacillus paralimentarius* (38). In our results, the type strains of *L. kimchii* and *L. bobalius* were in the same cluster, with a minimum CNI of 94.4%, while the *L. paralimentarius* type strain was present in a different cluster. The two clusters showed a maximum CNI value of only 93.9%. Thus, we can confirm that *L. bobalius* can be considered a later heterotypic synonym of *L. kimchii*, and we do not object to also considering both of them later heterotypic synonyms of *L. paralimentarius* since their clusters were extremely closely related. *Leuconostoc gelidum* (39), *Leuconostoc inhae* (40), and *Leuconostoc gasicomitatum* (41) were present in the same cluster, with a minimum CNI of 94.6%. The type strains of *Leuc. gasicomitatum* and *Leuc. inhae* even shared 99.5% CNI. Thus, *Leuc. inhae* should definitely be considered a later heterotypic synonym of *Leuc. gasicomitatum*. To our knowledge, this has not been suggested before. These two type strains shared 94.8% and 94.6% CNI with the *Leuc. gelidum* type strain, respectively, just on the border of being the same species.

In 2018, three *Lactobacillus gasseri* (42) strains were reclassified as *Lactobacillus paragasseri*, sp. nov. (43). We found both species in the same cluster, with a minimum CNI of 94.6%. Thus, this is a borderline case, and we do not object to *L. paragasseri* being considered a separate species. In a similar case, a subspecies of *Leuconostoc mesenteroides* (44) was recently introduced as the new species *Leuconostoc suionicum* (45). We found the type genomes of these species in the same cluster with a minimum CNI of 95.2%. This is again a borderline case.

*Lactobacillus casei* (46) and *Lactobacillus zeae* (47) were in the same cluster, with a minimum CNI of 95.4%. There has been much confusion surrounding the names *L. casei*, *L. paracasei*, and *L. zeae*. This is mainly the case because in 1971, strain ATCC 393 was chosen as the type strain for *L. casei* based on carbohydrate fermentation profiles (48). In 1980, this type strain was included in the Approved Lists of Bacterial Names (49).



However, the strain later appeared to be relatively distant to the other historical *L. casei* strains. Subsequent proposals to change this illogical type strain (e.g., reference 50) were rejected because they violated the International Code of Nomenclature of Bacteria (13). Thus, strain ATCC 393 remained the *L. casei* type strain, which had some confusing consequences. First, the name *L. zeae* was rejected because its type strain was too closely related to strain ATCC 393. In addition, the name *L. paracasei*, whose type strain was very closely related to the historical *L. casei* strains, was kept alive. The consequence was that many of the strains that were historically classified to *L. casei* should now be reclassified to *L. paracasei*, and that all strains historically classified to *L. zeae* should now be reclassified to *L. casei* (14). In this work, we confirmed that the *L. casei* type strain and the (rejected) *L. zeae* type strain are closely related, but not so closely that it would be absurd to consider them different species.

The species *Leuconostoc garlicum* has one high-quality genome in the NCBI database, but it was reclassified to *Leuconostoc lactis* (51) by our pipeline. The name *Leuc. garlicum* occurs in a few publications but has never been published, validly or otherwise. Therefore, we suggest that this species name should not be validated and that these strains should be classified to the species *Leuc. lactis*. Finally, we found that *Lactobacillus aviarius* subsp. *aviarius* and *Lactobacillus aviarius* subsp. *araffinosus* (52) had a CNI of 91.1%, and therefore, we suggest that the latter be renamed *Lactobacillus araffinosus* comb. nov. To our knowledge, it has never been suggested before that these should be separate species.

We labeled 13 genome clusters that contained no type genomes with a species name because their NCBI species label(s) matched their 16S rRNA hits. However, because the type strains of these species were not sequenced or their sequenced genomes were not of sufficient quality, we could not identify these *de novo* species with 100% certainty. Therefore, we propose that priority be given to (re)sequencing of these type strains. Further, we found that five species without NCBI species labels showed matches to type strains on the 16S rRNA level; these type strains should also be sequenced to be able to identify the *de novo* species with certainty. Finally, we discovered at least eight new species for which type strains should be chosen and deposited and species names should be picked.

To make it easy for people to classify their own (potential) LGC genomes against our *de novo* species taxonomy, we wrote a small classification tool. It is available at <https://github.com/SWittouck/proclasp> and includes a tutorial with an LGC reference database.

**Conclusions.** We constructed a genome-based species taxonomy of the *Lactobacillus* genus complex by performing single-linkage clustering on 2,459 high-quality genomes with a 94% core nucleotide identity cutoff. We found that the resulting species were discontinuous, fully exclusive, and almost fully transitive. On the basis of a comparison of the *de novo* species with published species, we proposed nine mergers of two or more species and one split of a species. Further, we discovered at least eight yet-to-be-named species that have not been published before, validly or otherwise. Finally, we have shown that our newly developed method to extract single-copy core genes allows for taxonomic studies on thousands of genomes on a desktop computer and is therefore applicable to other genera or even larger taxa. We believe that correcting the published species in the direction of our *de novo* taxonomy will lead to more meaningful species that show a consistent diversity.

## MATERIALS AND METHODS

**Downloading of genomes and gene prediction.** All genome assemblies classified to the *Lactobacillaceae* family (NCBI taxonomy id 33958) or *Leuconostocaceae* family (NCBI taxonomy id 81850) were downloaded from GenBank using the script `GenBank_get_genomes_by_taxon.py` that is part of the tool `pyani`, version 0.2.7 (53). Gene prediction was performed on all of these assemblies using Prodigal version 2.6.3 (54).

**Extraction of single-copy core genes.** To be able to rapidly extract single-copy core genes (SCGs) from our large genome data set, we devised a new approach based on seed genomes. Our strategy can be divided into four steps. In the first step, *n* seed genomes were selected completely at random, and their genes were clustered into gene families. The software program OrthoFinder (55) was used for gene



family inference; it performs all-versus-all blastp (56) search, normalizes the scores for sequence length and evolutionary distances between the genomes, applies a score cutoff per gene based on its best bidirectional hits, and finally clusters the resulting weighted graph using MCL (57). This process resulted in gene families for the seed genomes only. In the second step, candidate SCGs were selected from all seed gene families by identifying gene families present in more than  $k$  seed genomes. The third step was the identification of these candidate SCGs in all genomes. For this purpose, the seed sequences for each candidate SCG were first aligned using MAFFT (58). Next, profile HMMs were constructed from those alignments using the hmmbuild tool of HMMER (hmmer.org). All genomes were then scanned for occurrences of the candidate SCGs with the hmmsearch tool of HMMER. Finally, a homology score cutoff was determined for each profile HMM (see the next paragraph for the details), and these cutoffs were applied to the HMMER hits to yield all “real” hits of the candidate SCGs in all genomes. In the fourth step, the final SCGs were determined by retaining only candidate SCGs with exactly one gene (copy) in at least  $p$  percentage of all genomes.

For the training of profile-specific hmmer score cutoffs, the following strategy was used. First, if a gene was a hit of multiple profiles (multiple candidate SCGs), only the highest scoring profile was kept. Next, profile-specific cutoffs were determined as follows. Imagine we are selecting a score cutoff for a given profile. For each genome that has at least one gene among the hits of the profile, the best-scoring gene is considered a “true” hit. All other genes (second, third, etc., best hits of genomes) are considered “false” hits. A score threshold is then determined that maximizes the F-measure of the hits; this is the harmonic mean of the precision and recall. In effect, this means that the threshold is set in such a way that as many as possible “best hits of genomes” are included in the results, while at the same time as many as possible results should be a “best hit of a genome.” The optimization was performed using the R package ROCR (59). This approach was applied separately to every profile HMM (each one representing a candidate SCG).

The SCG extraction strategy we described here was implemented in progenomics: a general, in-development toolkit for prokaryotic comparative genomics ([github.com/SWittouck/progenomics](https://github.com/SWittouck/progenomics)). The parameter values used in this study were 30 for  $n$  (number of seed genomes), 25 for  $k$  (required number of seed genomes a candidate SCG should be present in), and 95 for  $p$  (required percentage of total genomes a final SCG should be present in). The software versions used were progenomics version 0.1.0, OrthoFinder version 2.1.2, blast version 2.6.0, MCL version 14-137, MAFFT version 7.407, HMMER version 3.1b2, and ROCR version 1.0.7.

**Genome quality control.** For each genome, two quality measures were calculated based on the SCGs: the completeness (percentage of SCGs present in the genome) and redundancy (percentage of SCGs having more than one copy in the genome). Only genomes with  $>90\%$  completeness and  $<10\%$  redundancy were retained. To compare our quality estimates to those available on the Genome Taxonomy Database (GTDB), we downloaded the metadata of the bacterial genomes of the latest GTDB release on 16 July 2019.

**Calculation of pairwise genome similarities.** For every unique pair of genomes, two similarity measures were computed: the core nucleotide identity (CNI) and average nucleotide identity (ANI). For the CNIs, multiple alignment was first performed for each SCG on the amino acid level using MAFFT version 7.407 (58) with the default parameters. Those amino acid alignments were then used as a reference to align the SCGs on the nucleotide level (reverse alignment). The aligned nucleotides were then concatenated into one large alignment (“supermatrix”). Next, pairwise nucleotide differences were calculated from this alignment using the distmat tool of EMBOSS version 6.6.0.0 (60), without correcting for multiple substitutions. Those differences were converted to similarities by subtracting them from 1. Pairwise ANIs were calculated using the tool fastANI (9), version 1.1, which quickly computes a relatively accurate approximation of ANI.

**Clustering of genomes.** Genomes were clustered into *de novo* species by joining together all genome pairs with a CNI similarity higher than 0.94. One could think of this as “nonhierarchical single-linkage clustering.” A simple algorithm to perform this clustering was implemented in R version 3.5.1 (61).

**Identification of type genomes.** To identify genomes of type strains (“type genomes”), we first downloaded assembly metadata for all of our genomes from the NCBI database using the rentrez R package version 1.2.1 (62). One of those metadata fields was the strain name of the assembled genome. Next, we compiled a list of all type strain names of validly published species of the six genera that make up the LGC. Since type strains are deposited in at least two but often many more culture collections, all of them have at least two names. To collect those type strain names, we developed an R package called tidytypes ([github.com/SWittouck/tidytypes](https://github.com/SWittouck/tidytypes), version 0.1.0), which searches the websites LPSN (15, 16), PNU (17), and StrainInfo (63). StrainInfo was used only to find extra synonyms of type strains of species and subspecies found on LPSN and PNU. By looking up those type strain names (including their many synonyms) in the assembly metadata from NCBI (which includes a strain name field), we were able to identify type genomes in our data set on a large scale and in an automated manner. We then supplemented this list by manually identifying extra type genomes in the following way. We gathered a list of species names that were present in the NCBI assembly metadata but for which we had not already found a type genome. For each of those species, we looked up the type strain names in the original paper that described the species. We then inspected the assembly metadata of the genomes classified to that species on NCBI in an attempt to find type genomes that were missed by the automated approach (for example because the species were not validly published). The result of the automated and manual approaches was a complete list of type genomes in our genome data set.

**Assignment of species names to genome clusters.** Two different strategies were employed to assign species names to the genome clusters: one based on the type genomes present in the clusters and one based on comparisons to 16S rRNA gene sequences of type strains. First, we used the list of type genomes to assign species names to genome clusters. If the cluster contained type genomes from exactly one species, we assigned this species name to the cluster. If it contained type genomes from more than one species, we concluded that those species were closely related enough to be considered heterotypic synonyms and assigned the oldest species name to the cluster. If two clusters contained type genomes from different subspecies of the same species, we proposed upgrading the cluster with the nontype subspecies (e.g., *Lactobacillus casei* subsp. *tolerans*) to species status (e.g., *Lactobacillus tolerans*).

For the genome clusters that did not contain one or more type genomes, we predicted 16S rRNA genes using barrnap version 0.9 (64). For all species and subspecies names on LPSN for which we did not find a type genome in our data set, we downloaded a 16S rRNA sequence from GenBank. We then scanned the 16S sequences extracted from our genomes against this reference 16S database using the blastn tool of blast version 2.6.0 (56). We then applied a percentage identity cutoff of 98% on those hits. When a genome cluster contained one or more genomes with a match to an official species through the 16S approach, and the same species was listed in the species labels of one or more genomes, we assigned this species name to the genome cluster as a “best guess.” When a genome cluster yielded 16S genes but none of them showed a match to the reference database, we considered it a new species. Finally, some genome clusters remained that could not be identified or be considered a new species; we labeled those “unidentified.”

**Inference of species phylogeny.** For each genome cluster, the genome with the largest number of SCGs present was selected as a representative genome. We then selected the 100 SCGs that had the largest single-copy presence counts in those representative genomes. We then constructed a supermatrix with the selected genomes and SCGs, in the same way as described above in the section “Calculation of pairwise genome similarities” (including reverse alignment). Next, we trimmed columns that had gaps in more than 1% of the genomes using trimAl version 1.4.rev15 (65). We inferred a maximum likelihood tree using RAxML version 8.2.11 (66) using the general time reversible model for nucleotide substitutions and the CAT method for modeling mutation rate heterogeneity across columns. We used the “-f a” option of RAxML, which combines rapid bootstrapping (67) with a slow search of the tree space starting from the bootstrap trees. The number of bootstrap trees was determined by the “N autoMRE” option, which will stop looking for bootstrap trees when they do not seem to add any extra information (68).

**Data processing and visualization.** All processing and visualization of table-format data were done in R version 3.5.1 (61) using the tidyverse set of packages, version 1.2.1 (69). Phylogenetic tree visualization and annotation were done using ggtree version 1.12.7 (70).

**Availability of data and material.** NCBI assembly accession numbers for the genomes used in this study can be found in Table S1 in the supplemental material. The code used for this study was split over four repositories. The main pipeline (called “legen,” for Lactobacillus Evolutionary Genomics) is available at [https://github.com/SWittouck/legen\\_pipeline](https://github.com/SWittouck/legen_pipeline) (version 3.1). R Markdown scripts for the data analysis, including the creation of all figures and tables, are available at [https://github.com/SWittouck/legen\\_data\\_analysis](https://github.com/SWittouck/legen_data_analysis) (version 3.0). This repository also contains the necessary data files and can thus be run on its own. Some of the steps of the pipeline, including the extraction of single-copy core genes, were implemented in the progenomics toolkit for prokaryotic comparative genomics for maximal reusability. This toolkit is available at <https://github.com/SWittouck/progenomics> (version 0.1.0). Finally, searching taxonomic websites to obtain type strain names of published prokaryotic species was implemented in an R package called tidytypes, which can be found at <https://github.com/SWittouck/tidytypes> (version 0.1.0).

## SUPPLEMENTAL MATERIAL

Supplemental material for this article may be found at <https://doi.org/10.1128/mSystems.00264-19>.

**FIG S1**, EPS file, 0.2 MB.

**FIG S2**, EPS file, 0.04 MB.

**FIG S3**, EPS file, 0.1 MB.

**TABLE S1**, CSV file, 0.3 MB.

**TABLE S2**, CSV file, 0.03 MB.

**TABLE S3**, CSV file, 0.02 MB.

## ACKNOWLEDGMENTS

We thank Camille Allonsius for proofreading the text for readability and clarity. In addition, we want to thank the entire Lab of Applied Microbiology and Biotechnology at the University of Antwerp and the Computational Systems Biology group at the KU Leuven for providing feedback on the general ideas and results. We also acknowledge Bruno Pot for his advice on taxonomic databases and Giovanna Felis for her constructive feedback. Finally, we want to acknowledge Rhambo (R.I.P.) and LAMBorGENEi for performing the computations.

This study was supported by the Research Foundation Flanders (grant 11A0618N),

the Flanders Innovation and Entrepreneurship Agency (grants IWT-SB 141198 and IWT/50052), and the University of Antwerp (grant FFB150344).

We declare that we have no competing interests.

S. Wittouck and S. Lebeer designed the study. S. Wittouck performed the analyses, while S. Wuyts, C. J. Meehan, and V. van Noort provided important input on the bioinformatics. S. Wittouck and S. Lebeer interpreted the results. S. Wittouck wrote the draft article. All authors provided feedback on the structure and writing of the article. All authors read and approved the final article.

## REFERENCES

- Sun Z, Harris HMB, McCann A, Guo C, Argimón S, Zhang W, Yang X, Jeffery IB, Cooney JC, Kagawa TF, Liu W, Song Y, Salvetti E, Wrobel A, Rasinkangas P, Parkhill J, Rea MC, O'Sullivan O, Ritari J, Douillard FP, Ross RP, Yang R, Briner AE, Felis GE, de Vos WM, Barrangou R, Klaenhammer TR, Caufield PW, Cui Y, Zhang H, O'Toole PW. 2015. Expanding the biotechnology potential of lactobacilli through comparative genomics of 213 strains and associated genera. *Nat Commun* 6:8322. <https://doi.org/10.1038/ncomms9322>.
- Duar RM, Lin XB, Zheng J, Martino ME, Grenier T, Perez-Munoz M, Leulier F, Ganzle MG, Walter J. 2017. Lifestyles in transition: evolution and natural history of the genus *Lactobacillus*. *FEMS Microbiol Rev* 41(Suppl 1):S27–S48. <https://doi.org/10.1093/femsre/fux030>.
- Heeney DD, Gareau MG, Marco ML. 2018. Intestinal *Lactobacillus* in health and disease, a driver or just along for the ride? *Curr Opin Biotechnol* 49:140–147. <https://doi.org/10.1016/j.copbio.2017.08.004>.
- Petrova MI, Lievens E, Malik S, Imholz N, Lebeer S. 2015. *Lactobacillus* species as biomarkers and agents that can promote various aspects of vaginal health. *Front Physiol* 6:81. <https://doi.org/10.3389/fphys.2015.00081>.
- Wittouck S, Wuyts S, Lebeer S. 2019. Towards a genome-based reclassification of the genus *Lactobacillus*. *Appl Environ Microbiol* 85:e02155–18. <https://doi.org/10.1128/AEM.02155-18>.
- Zheng J, Ruan L, Sun M, Ganzle M. 2015. A genomic view of lactobacilli and pediococci demonstrates that phylogeny matches ecology and physiology. *Appl Environ Microbiol* 81:7233–7243. <https://doi.org/10.1128/AEM.02116-15>.
- Salvetti E, Harris HMB, Felis GE, O'Toole PW. 2018. Comparative genomics of the genus *Lactobacillus* reveals robust phylogroups that provide the basis for reclassification. *Appl Environ Microbiol* 84:e00993–18. <https://doi.org/10.1128/AEM.00993-18>.
- Parks DH, Chuvochina M, Waite DW, Rinke C, Skarshewski A, Chaumeil P-A, Hugenholtz P. 2018. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat Biotechnol* 36:996–1004. <https://doi.org/10.1038/nbt.4229>.
- Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. 2018. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat Commun* 9:5114. <https://doi.org/10.1038/s41467-018-07641-9>.
- Wright ES, Baum DA. 2018. Exclusivity offers a sound yet practical species criterion for bacteria despite abundant gene flow. *BMC Genomics* 19:724. <https://doi.org/10.1186/s12864-018-5099-6>.
- Konstantinidis KT, Tiedje JM. 2005. Genomic insights that advance the species definition for prokaryotes. *Proc Natl Acad Sci U S A* 102:2567–2572. <https://doi.org/10.1073/pnas.0409727102>.
- Richter M, Rosselló-Móra R. 2009. Shifting the genomic gold standard for the prokaryotic species definition. *Proc Natl Acad Sci U S A* 106:19126–19131. <https://doi.org/10.1073/pnas.0906412106>.
- Judicial Commission of the International Committee on Systematics of Bacteria. 2008. The type strain of *Lactobacillus casei* is ATCC 393, ATCC 334 cannot serve as the type because it represents a different taxon, the name *Lactobacillus paracasei* and its subspecies names are not rejected and the revival of the name "*Lactobacillus zeae*" contravenes Rules 51b (1) and (2) of the International Code of Nomenclature of Bacteria. Opinion 82. *Int J Syst Evol Microbiol* 58:1764–1765. <https://doi.org/10.1099/ijs.0.2008/005330-0>.
- Wuyts S, Wittouck S, De Boeck I, Allonsius CN, Pasolli E, Segata N, Lebeer S. 2017. Large-scale phylogenomics of the *Lactobacillus casei* group highlights taxonomic inconsistencies and reveals novel clade-associated features. *mSystems* 2:e00061-17. <https://doi.org/10.1128/mSystems.00061-17>.
- Euzéby JP. 1997. List of Bacterial Names with Standing in Nomenclature: a folder available on the Internet. *Int J Syst Bacteriol* 47:590–592. <https://doi.org/10.1099/00207713-47-2-590>.
- Parte AC. 2014. LPSN—list of prokaryotic names with standing in nomenclature. *Nucleic Acids Res* 42:D613–D616. <https://doi.org/10.1093/nar/gkt111>.
- DSMZ. 2017. Prokaryotic Nomenclature Up-to-Date. Deutsche Sammlung von Mikroorganismen und Zellkulturen, Braunschweig, Germany. Accessed 30 January 2019.
- Heo J, Saitou S, Tamura T, Cho H, Kim J-S, Joa J-H, Kim J-S, Kwon S-W, Kim S-J. 2018. *Lactobacillus nuruki* sp. nov., isolated from Nuruk, a Korean fermentation starter. *Int J Syst Evol Microbiol* 68:3273–3278. <https://doi.org/10.1099/ijsem.0.002976>.
- Douillard FP, Ribbera A, Kant R, Pietilä TE, Järvinen HM, Messing M, Randazzo CL, Paulin L, Laine P, Ritari J, Caggia C, Lähteinen T, Brouns SJJ, Satokari R, von Ossowski I, Reunanen J, Palva A, de Vos WM. 2013. Comparative genomic and functional analysis of 100 *Lactobacillus rhamnosus* strains and their comparison with strain GG. *PLoS Genet* 9:e1003683. <https://doi.org/10.1371/journal.pgen.1003683>.
- Martino ME, Bayjanov JR, Caffrey BE, Wels M, Joncour P, Hughes S, Gillet B, Kleerebezem M, van Hijum S, Leulier F. 2016. Nomadic lifestyle of *Lactobacillus plantarum* revealed by comparative genomics of 54 strains isolated from different habitats. *Environ Microbiol* 18:4974–4989. <https://doi.org/10.1111/1462-2920.13455>.
- Dykhuizen DE, Green L. 1991. Recombination in *Escherichia coli* and the definition of biological species. *J Bacteriol* 173:7257–7268. <https://doi.org/10.1128/jb.173.22.7257-7268.1991>.
- Cohan FM, Kane M. 2001. Bacterial species and speciation. *Syst Biol* 50:513–524. <https://doi.org/10.1080/10635150118398>.
- McFrederick QS, Vuong HQ, Rothman JA. 2018. *Lactobacillus micheneri* sp. nov., *Lactobacillus timberlakei* sp. nov. and *Lactobacillus quenuiae* sp. nov., lactic acid bacteria isolated from wild bees and flowers. *Int J Syst Evol Microbiol* 68:1879–1884. <https://doi.org/10.1099/ijsem.0.002758>.
- Tanasupawat S, Shida O, Okada S, Komagata K. 2000. *Lactobacillus acidipiscis* sp. nov. and *Weissella thailandensis* sp. nov., isolated from fermented fish in Thailand. *Int J Syst Evol Microbiol* 50:1479–1485. <https://doi.org/10.1099/00207713-50-4-1479>.
- Lee S-H, Ku H-J, Ahn M-J, Hong J-S, Lee SH, Shin H, Lee KC, Lee J-S, Ryu S, Jeon CO, Lee J-H. 2015. *Weissella jogaejeotgali* sp. nov., isolated from jogae jeotgal, a traditional Korean fermented seafood. *Int J Syst Evol Microbiol* 65:4674–4681. <https://doi.org/10.1099/ijsem.0.000631>.
- Charlton DB, Nelson ME, Werkman CH. 1934. Physiology of *Lactobacillus fructivorans* sp. nov. isolated from spoiled salad dressing. *Iowa State Coll J Sci* 9:1–11.
- Kitahara K, Kaneko T, Goto O. 1957. Taxonomic studies on the hiochi-bacteria, specific saprophytes of sake. *J Gen Appl Microbiol* 3:111–120. <https://doi.org/10.2323/jgam.3.111>.
- Suzuki K, Asano S, Iijima K, Kitamoto K. 2008. Sake and beer spoilage lactic acid bacteria - a review. *J Inst Brewing* 114:209. <https://doi.org/10.1002/j.2050-0416.2008.tb00331.x>.
- Parker CT, Tindall BJ, Garrity GM. 2019. International code of nomenclature of prokaryotes. *Int J Syst Evol Microbiol* 69:51–111. <https://doi.org/10.1099/ijsem.0.000778>.
- Lindner P. 1887. Über ein neues in Malzmaischen vorkommendes, milchsäurebildendes. *Ferment Wochenschr Brauerei* 4:437–440.
- Dai K, Nishizaki Y, Fujino Y, Ohshima T, Ohmomo S, Ogata S. 2009. *Pediococcus lolii* sp. nov., isolated from ryegrass silage. *Int J Syst Evol Microbiol* 59:1007–1010. <https://doi.org/10.1099/ijs.0.005793-0>.
- Wieme A, Cleenwerck I, Van Landschoot A, Vandamme P. 2012. *Pedio-*

- coccus lolii DSM 19927T and JCM 15055T are strains of *Pediococcus acidilactici*. Int J Syst Evol Microbiol 62:3105–3108. <https://doi.org/10.1099/ijs.0.046201-0>.
33. Naser SM, Vancanneyt M, Snaauwaert C, Vrancken G, Hoste B, De Vuyst L, Swings J. 2006. Reclassification of *Lactobacillus amylophilus* LMG 11400 and NRRL B-4435 as *Lactobacillus amylophilus* sp. nov. Int J Syst Evol Microbiol 56:2523–2527. <https://doi.org/10.1099/ijs.0.64463-0>.
  34. Nakamura LK. 1981. *Lactobacillus amylovorus*, a new starch-hydrolyzing species from cattle waste-corn fermentations. Int J Syst Evol Microbiol 31:56–63. <https://doi.org/10.1099/00207713-31-1-56>.
  35. Pang H, Kitahara M, Tan Z, Wang Y, Qin G, Ohkuma M, Cai Y. 2012. Reclassification of *Lactobacillus kimchii* and *Lactobacillus bobalius* as later subjective synonyms of *Lactobacillus paralimentarius*. Int J Syst Evol Microbiol 62:2383–2387. <https://doi.org/10.1099/ijs.0.035329-0>.
  36. Yoon JH, Kang SS, Mheen TI, Ahn JS, Lee HJ, Kim TK, Park CS, Kho YH, Kang KH, Park YH. 2000. *Lactobacillus kimchii* sp. nov., a new species from kimchi. Int J Syst Evol Microbiol 50:1789–1795. <https://doi.org/10.1099/00207713-50-5-1789>.
  37. Manes-Lazaro R, Ferrer S, Rodas AM, Urdiain M, Pardo I. 2008. *Lactobacillus bobalius* sp. nov., a lactic acid bacterium isolated from Spanish Bobal grape must. Int J Syst Evol Microbiol 58:2699–2703. <https://doi.org/10.1099/ijs.0.65695-0>.
  38. Cai Y, Okada H, Mori H, Benno Y, Nakase T. 1999. *Lactobacillus paralimentarius* sp. nov., isolated from sourdough. Int J Syst Evol Microbiol 49:1451–1455. <https://doi.org/10.1099/00207713-49-4-1451>.
  39. Shaw BG, Harding CD. 1989. *Leuconostoc gelidum* sp. nov. and *Leuconostoc carnosum* sp. nov. from chill-stored meats. Int J Syst Evol Microbiol 39:217–223. <https://doi.org/10.1099/00207713-39-3-217>.
  40. Kim B, Lee J, Jang J, Kim J, Han H. 2003. *Leuconostoc inhae* sp. nov., a lactic acid bacterium isolated from kimchi. Int J Syst Evol Microbiol 53:1123–1126. <https://doi.org/10.1099/ijs.0.02463-0>.
  41. Björkroth KJ, Geisen R, Schillinger U, Weiss N, De Vos P, Holzapfel WH, Korkeala HJ, Vandamme P. 2000. Characterization of *Leuconostoc gasi-comitatum* sp. nov., associated with spoiled raw tomato-marinated broiler meat strips packaged under modified-atmosphere conditions. Appl Environ Microbiol 66:3764–3772. <https://doi.org/10.1128/AEM.66.9.3764-3772.2000>.
  42. Lauer E, Kandler O. 1980. *Lactobacillus gasseri* sp. nov., a new species of the subgenus *Thermobacterium*. Zentralbl. Bakteriologie I Abt. Orig. C 1:75–78. [https://doi.org/10.1016/S0172-5564\(80\)80019-4](https://doi.org/10.1016/S0172-5564(80)80019-4).
  43. Tanizawa Y, Tada I, Kobayashi H, Endo A, Maeno S, Toyoda A, Arita M, Nakamura Y, Sakamoto M, Ohkuma M, Tohno M. 2018. *Lactobacillus paragasseri* sp. nov., a sister taxon of *Lactobacillus gasseri*, based on whole-genome sequence analyses. Int J Syst Evol Microbiol 68:3512–3517. <https://doi.org/10.1099/ijsem.0.003020>.
  44. Tsenkovskii L. 1878. Gel formation of sugar beet solutions. Proc Soc Nat Sci Imperial Univ Kharkov 12:137–167. (In Russian.)
  45. Jeon HH, Kim KH, Chun BH, Ryu BH, Han NS, Jeon CO. 2017. A proposal of *Leuconostoc mesenteroides* subsp. *jonggajibkimchii* subsp. nov. and reclassification of *Leuconostoc suionicum* sp. nov. based on complete genome sequences. Int J Syst Evol Microbiol 67:2225–2230. <https://doi.org/10.1099/ijsem.0.001930>.
  46. Orla-Jensen S. 1916. Mælkeri-bakteriologi. Schønberg'ske Forlag, Copenhagen, Denmark. (In Danish.)
  47. Kuznetsov VD. 1959. A new species of lactic acid bacteria. Mikrobiologiya 28:248–251.
  48. Hansen PA, Lessel EF. 1971. *Lactobacillus casei* (Orla-Jensen) comb. nov. Int J Syst Bacteriol 21:69–71. <https://doi.org/10.1099/00207713-21-1-69>.
  49. Skerman VBD, McGowan V, Sneath P (ed). 1980. Approved Lists of Bacterial Names. Int J Syst Bacteriol 30:225–420. <https://doi.org/10.1099/00207713-30-1-225>.
  50. Dellaglio F, Dicks LMT, Du Toit M, Torriani S. 1991. Designation of ATCC 334 in place of ATCC 393 (NCDO 161) as the neotype strain of *Lactobacillus casei* subsp. *casei* and rejection of the name *Lactobacillus paracasei* (Collins et al., 1989): request for an opinion. Int J Syst Bacteriol 41:340–342. <https://doi.org/10.1099/00207713-41-2-340>.
  51. Garvie EI. 1960. The genus *Leuconostoc* and its nomenclature. J Dairy Res 27:283–292. <https://doi.org/10.1017/S0022029900010359>.
  52. Fujisawa T, Shirasaka S, Watabe J, Mitsuoka T. 1984. *Lactobacillus aviarius* sp. nov.: a new species isolated from the intestine of chickens. Syst Appl Microbiol 5:414–420. [https://doi.org/10.1016/S0723-2020\(84\)80042-9](https://doi.org/10.1016/S0723-2020(84)80042-9).
  53. Pritchard L, Glover RH, Humphris S, Elphinstone JG, Toth IK. 2016. Genomics and taxonomy in diagnostics for food security: soft-rotting enterobacterial plant pathogens. Anal Methods 8:12–24. <https://doi.org/10.1039/C5AY02550H>.
  54. Hyatt D, Chen G-L, Locascio PF, Land ML, Larimer FW, Hauser LJ. 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 11:119. <https://doi.org/10.1186/1471-2105-11-119>.
  55. Emms DM, Kelly S. 2015. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol 16:157. <https://doi.org/10.1186/s13059-015-0721-2>.
  56. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol 215:403–410. [https://doi.org/10.1016/S0022-2836\(05\)80360-2](https://doi.org/10.1016/S0022-2836(05)80360-2).
  57. Dongen S. 2000. A cluster algorithm for graphs. Centre for Mathematics and Computer Science (CWI), Amsterdam, The Netherlands.
  58. Katoh K, Misawa K, Kuma K-I, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res 30:3059–3066. <https://doi.org/10.1093/nar/gkf436>.
  59. Sing T, Sander O, Beerenwinkel N, Lengauer T. 2005. ROCr: visualizing classifier performance in R. Bioinformatics 21:3940–3941. <https://doi.org/10.1093/bioinformatics/bti623>.
  60. Rice P, Longden I, Bleasby A. 2000. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet 16:276–277. [https://doi.org/10.1016/S0168-9525\(00\)00204-2](https://doi.org/10.1016/S0168-9525(00)00204-2).
  61. R Core Team. 2018. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
  62. Winter DJ. 2017. rentrez: an R package for the NCBI eUtils API. PeerJ Preprints 5:e3179v2. <https://doi.org/10.7287/peerj.preprints.3179v2>.
  63. Verslyppe B, De Smet W, De Baets B, De Vos P, Dawyndt P. 2014. StrainInfo introduces electronic passports for microorganisms. Syst Appl Microbiol 37:42–50. <https://doi.org/10.1016/j.syapm.2013.11.002>.
  64. Seemann T. 2018. barrnap 0.9: rapid ribosomal RNA prediction.
  65. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25:1972–1973.
  66. Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313. <https://doi.org/10.1093/bioinformatics/btu033>.
  67. Stamatakis A, Hoover P, Rougemont J. 2008. A rapid bootstrap algorithm for the RAxML Web servers. Syst Biol 57:758–771. <https://doi.org/10.1080/10635150802429642>.
  68. Pattengale N, Alipour M, Bininda-Emonds O, Moret B, Stamatakis A. 2009. How many bootstrap replicates are necessary? J Comput Biol 5541:184–200.
  69. Wickham H. 2017. tidyverse: easily install and load the “Tidyverse.”
  70. Yu G, Smith DK, Zhu H, Guan Y, Lam T. 2016. ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods Ecol Evol 8:28–36. <https://doi.org/10.1111/2041-210X.12628>.